Wide-Band Audio Coding based on Frequency Domain Linear Prediction

Petr Motlicek
Idiap Research Institute, Martigny, Switzerland, email: motlicek@idiap.ch
Sriram Ganapathy
Johns Hopkins University, Baltimore, USA, email: ganapathy@jhu.edu
Hynek Hermansky
Johns Hopkins University, Baltimore, USA, email: hynek@jhu.edu
Harinath Garudadri
Qualcomm Inc., San Diego, USA, email: hgarudad@qualcomm.com

***Abstract*—In this paper, we re-visit an original concept of speech coding in which the signal is separated into the carrier modulated by the signal envelope. A recently developed technique, called frequency domain linear prediction (FDLP), is applied for the efficient estimation of the envelope. The processing in the temporal domain allows for a straightforward emulation of the forward temporal masking. This, combined with an efficient non-uniform sub-band decomposition and application of noise shaping in spectral domain instead of temporal domain (a technique to suppress artifacts in tonal audio signals), yields a codec that does not rely on the linear speech production model but rather uses well accepted concept of frequency-selective auditory perception. As such, the codec is not only specific for coding speech but also well suited for coding other important acoustic signals such as music and mixed content. The quality of the proposed codec at 66 kbps is evaluated using objective and subjective quality assessments. The evaluation indicates competitive performance with the MPEG codecs operating at similar bit-rates.**

***Index Terms*—Speech coding, audio coding, frequency domain linear prediction (FDLP), perceptual evaluation of audio quality (PEAQ).**

I. Introduction

Modern speech coding algorithms are based on source-filter model, wherein the model parameters are extracted using linear prediction principles applied in temporal domain [1]. Most popular audio coding algorithms are based on exploiting psychoacoustic models in the spectral domain [2], [3]. In this work, we explore signal processing methods to code speech and audio signals in a unified approach.

In traditional applications of speech coding (i.e., for conversational services), the algorithmic delay of the codec is one of the most critical variables. However, there are many services, such as downloading the audio files, voice messaging, or push-to-talk communications, where the issue of the codec delay is much less critical. This allows for a whole set of different coding techniques that could be more effective than the conventional short-term frame based coding techniques.

Due to the development of new audio services, there has been a need for new audio compression techniques which would provide sufficient generality, i.e., the ability to encode any kind of the input signal (speech, music, signals with mixed audio sources, transient signals). Traditional approaches to speech coding based on source-filter model have become very successful in commercial applications for toll quality conversational services. However, they do not perform well for mixed signals in many multimedia services. On the other hand, perceptual codecs have become useful in media coding applications, but are not as efficient for speech content. These contradictions have recently turned into new initiatives of standardization organizations (3GPP, ITU-T, and MPEG), which are interested in developing a codec for compressing mixed signals, e.g. speech and audio content.

This paper describes a coding technique that re-visits (similar to [4]), the original concept of the first speech coder [5], where the speech is seen as a carrier signal modulated by its temporal envelope. Our approach (first introduced in [6]) differs from [4] in use of frequency domain linear prediction (FDLP) [7], [8], [9], [10], that allows for the approximation of temporal (Hilbert) envelopes of sub-band energies by an auto-regressive (AR) model. Unlike temporal noise shaping (TNS) [7], which also uses FDLP and forms a part of the MPEG-2/4 AAC codec, where FDLP is applied to solve problems with transient attacks (impulses), the proposed codec employs FDLP to approximate relatively long (hundreds of milliseconds) segments of Hilbert envelopes in individual frequency sub-bands. Another approach, described in [11], exploits FDLP for sinusoidal audio coding using short-term segments.

The goal is to develop a novel wide-band (WB)-FDLP audio coding system that would explore new potentials in encoding mixed input including speech and audio by taking into account relatively long acoustic context directly in the initial step of encoding the input signal. Due to this acoustic context, the proposed coding technique is intended to be exploited in non-interactive audio services. Unlike interactive audio services such as VoIP or interactive games, the real-time constraints for the proposed codec are not stringent.

The paper is organized as follows: Section II discusses fundamental aspects of the FDLP technique. Section III mentions initial attempts to exploit FDLP for narrow-band speech coding. Section IV describes the WB-FDLP audio codec in general and Section V gives the detailed description of the

major blocks in the codec. Objective quality evaluations of the individual blocks are given in Section VI. Section VII provides subjective quality assessment of the proposed codec compared with state-of-the-art MPEG audio codecs. Section VIII contains discussions and summarizes important aspects.

II. Frequency domain linear prediction (FDLP)

Inertia of the human vocal tract organs makes the modulations in the speech signal to vary gradually. While short-term predictability within time-spans of $10-20$ ms and AR modeling of the signal have been used effectively [12], [13], there exists a longer-term predictability due to inertia of human vocal organs and their neural control mechanisms. Therefore, the temporal evolutions of vocal tract shapes (and subsequently also of the short-term spectral envelopes of the signal) are predictable. In terms of compression efficiency, it is desirable to capitalize on this predictability by processing longer temporal context for coding rather than processing every short-term segments independently. While such an approach obviously introduces longer algorithmic delay, the efficiency gained may justify its deployment in many evolving communications applications. Initial encouraging experimental results were achieved on very low bit-rate speech coding [6], and feature extraction for automatic speech recognition [14].

In the proposed audio codec, we utilize the concept of linear prediction in spectral domain on sub-band signals. After decomposing the signal into the individual critical-bandwidth sub-bands, the sub-band signals are characterized by their envelope (amplitude) and carrier (phase) modulations. FDLP is then able to exploit the predictability of slowly varying amplitude modulations. Spectral representation of amplitude modulation in sub-bands, also called "Modulation Spectra", have been used in many engineering applications. Early work done in [15] for predicting speech intelligibility and characterizing room acoustics are now widely used in the industry [16]. Recently, there has been many applications of such concepts for robust speech recognition [17], [18], audio coding [4], noise suppression [19], etc. In order to use information in modulation spectrum (at important frequencies starting from low range), a signal over relatively long time scales has to be processed. This is also the case of FDLP.

Defining the analytic signal in the sub-band as $s_a(n) = s(n) + j\hat{s}(n)$, where $\hat{s}(n)$ is the Hilbert transform of $s(n)$, the Hilbert envelope of $s(n)$ is defined as $|s_a(n)|^2$ (squared magnitude of the analytic signal) and the phase is represented by instantaneous frequency, denoted by the first derivative of $\angle s_a(n)$ (scaled by $1/2\pi$). Often, the term Hilbert carrier denoted as $cos(\angle s_a(n))$ is used for representing the phase. Here, n denotes time samples.

As it will be shown in Section II-A, FDLP parameterizes the Hilbert envelope of the input signal $s(n)$. FDLP can be seen as a method analogous to temporal domain linear prediction (TDLP) [20]. In the case of TDLP, the AR model approximates the power spectrum of the input signal. The FDLP fits an AR model to the Hilbert envelope of the input signal. Using FDLP we can adaptively capture fine temporal details with high temporal resolution. At the same time, FDLP summarizes the temporal evolution of the signal over hundreds of milliseconds. Figure 1 shows an example of speech signal, its Hilbert envelope (obtained using the technique based on discrete Fourier transform [9]) and AR model estimated by FDLP.

A. Envelope Estimation

In this section, we describe, in detail, the approximation of temporal envelopes by AR model obtained using FDLP. To simplify the notation, we present the full-band version of the technique. The sub-band version is identical except that the technique is applied to the sub-band signal obtained by a filter bank decomposition.

In the previous section, we defined the input discrete time-domain sequence as $s(n)$ for time samples $n = 0, \ldots, N-1$, where N denotes the segment length. Its Fourier power spectrum $P(\omega_k)$ (sampled at discrete frequencies $\omega_k = \frac{2\pi}{N}k$; $k = 0, \ldots, N-1$) is given as

$$P(\omega_k) = |S(e^{j\omega_k})|^2, \tag{1}$$

where $S(e^{j\omega_k}) = Z\{s(n)\}|_{z=e^{j\omega_k}}$. $Z\{.\}$ stands for the z-transform. Let the notation $F\{.\}$ denote discrete Fourier transform (DFT) which is equivalent to z-transform with $z = e^{j\omega_k}$.

It has been shown, e.g., in [20], that the conventional TDLP fits the discrete power spectrum of an all-pole model $\hat{P}(\omega_k)$ to $P(\omega_k)$ of the input signal. Unlike TDLP, where the time-domain sequence $s(n)$ is modeled by linear prediction, FDLP applies linear prediction on the frequency-domain representation of the sequence. In our case, $s(n)$ is first transformed by discrete cosine transform (DCT). It can be shown that the DCT type I odd (DCT-Io) needs to be used [21]. DCT-Io can also be viewed as the symmetrical extension of $s(n)$ so that a new time-domain sequence $q(m)$ is obtained ($m = 0, \ldots, M-1$, and $M = 2N$) and then DFT projected (i.e., relationship between the DFT and the DCT-Io). We obtain the real-valued sequence $Q(\omega_k) = F\{q(m)\}$, where $k = 0, \ldots, M-1$. Process of symmetrical extension allows to avoid problems with continuity at boundaries of the time signal (often called Gibbs-type ringing).

We then estimate the frequency domain prediction error $E(\omega_k)$ as a linear combination of $Q(\omega_k)$ consisting of p real prediction coefficients b_i

$$E(\omega_k) = Q(\omega_k) - \sum_{i=1}^{p} b_i Q(\omega_k - \omega_i). \tag{2}$$

b_i are found so that the squared prediction error is minimized [20]. In the case of TDLP, minimizing the total error is equivalent to the minimization of the integrated ratio of the signal spectrum $P(\omega_k)$ to its model approximation $\hat{P}(\omega_k)$

$$E_{TDLP} \approx \frac{1}{N} \sum_{k=0}^{N-1} \frac{P(\omega_k)}{\hat{P}(\omega_k)}. \tag{3}$$

In the case of FDLP, we can interpret $Q(\omega_k)$ as a discrete, real, causal, stable sequence (consisting of frequency samples). Its discrete power spectrum will be estimated through the concept of discrete Hilbert transform relationships [22]. $Q(\omega_k)$ can be

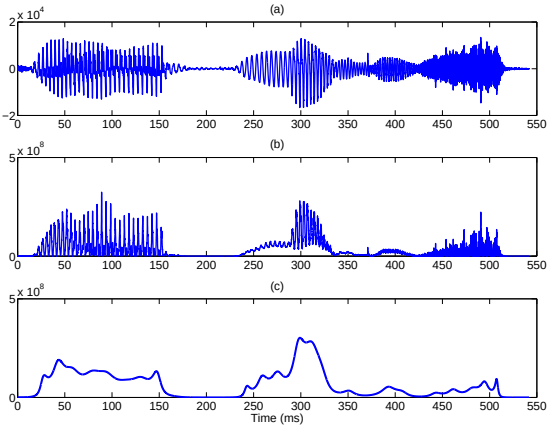


Fig. 1. Illustration of the AR modeling property of FDLP: (a) a portion of speech signal, (b) its Hilbert envelope, and (c) all-pole model obtained using FDLP.

expressed as the sum of $Q^e(\omega_k)$ and $Q^o(\omega_k)$, denoting an even sequence and an odd sequence, respectively; thus $Q(\omega_k) = Q^e(\omega_k) + Q^o(\omega_k)$. Its Fourier transform

$$\phi(m) = F\{Q(\omega_k)\} = \phi^R(m) + j\phi^I(m), \quad (4)$$

where R and I stand for real and imaginary parts of $\phi(m)$, respectively. It has been shown (e.g., [22]) that $\phi^R(m) = F\{Q^e(\omega_k)\}$ and $\phi^I(m) = F\{Q^o(\omega_k)\}$. By taking the Fourier transform of $Q^e(\omega_k)$, the original sequence $q(m)$ is obtained

$$F\{Q^e(\omega_k)\} = \phi^R(m) = Cq(m). \quad (5)$$

C stands for a constant. The relations between $F\{Q^e(\omega_k)\}$ and $F\{Q^o(\omega_k)\}$, called the Kramers-Kronig relations, are given by the discrete Hilbert transform (partial derivatives of real and imaginary parts of an analytic function [23]), thus

$$\phi(m) = \phi^R(m) + j\phi^I(m) = C\big(q(m) + jH\{q(m)\}\big), \quad (6)$$

where $H\{.\}$ stands for Hilbert transformation. $|\phi(m)|^2$ is called the Hilbert envelope (squared magnitude of the analytic signal $\phi(m)$). Prediction error is proportional to the integrated ratio of $|\phi(m)|^2$ and its FDLP approximation $A^2(m)$

$$E_{FDLP} \approx \frac{1}{M}\sum_{m=0}^{M-1}\frac{|\phi(m)|^2}{A^2(m)}. \quad (7)$$

$A^2(m)$ stands for squared magnitude frequency response of the all-pole model. Equation 7 can be interpreted in such a way that the FDLP all-pole model fits Hilbert envelope of the symmetrically extended time-domain sequence $s(n)$. FDLP models the time-domain envelope in the same way as TDLP models the spectral envelope. Therefore, the same properties appear, such as accurate modeling of peaks rather than dips.

Further, the Hilbert envelope $|\phi(m)|^2$ is available and can be modified (before applying linear prediction). Thus, e.g., compressing $|\phi(m)|^2$ by a root function $[.]^{\frac{1}{r}}$ turns Equation 7 into

$$E_{FDLP} \approx \frac{1}{M}\sum_{m=0}^{M-1}\frac{|\phi(m)|^{\frac{2}{r}}}{A^{\frac{2}{r}}(m)}. \quad (8)$$

As a consequence, the new model will fit dips more accurately than the original model. This technique has been proposed for TDLP (called spectral transform linear prediction (STLP) [24]), and we apply this scheme for FDLP.

III. FDLP FOR NARROW-BAND SPEECH CODING

Initial experiments aiming at narrow-band (NB) speech coding (8 kHz), reported in [6], suggest that FDLP applied on long temporal segments and excited with white noise signal provides a highly intelligible speech, but with whisper-like quality without any voicing at bit-rates below 1 kbps. In these experiments, the input speech was split into non-overlapping segments (hundreds of milliseconds long). Then, each segment was processed by DCT and partitioned into unequal frequency sub-segments to obtain critical band-sized sub-bands. FDLP approximation was applied on each sub-band by carrying out auto-correlation linear prediction (LP) analysis on the sub-segments of DCT transformed signals, yielding line spectral pair (LSP) descriptors of FDLP models. Resulting AR models approximate the Hilbert envelopes in critical band-sized sub-bands.

In case of very low bit-rate speech coding ($\sim$ 1 kbps), a frequency decomposition into 15 sub-bands was performed for every 1000ms long input segment. In each sub-band, the FDLP model of order of 20 was estimated. FDLP sub-band residuals (these signals represent sub-band Hilbert carriers for the sub-band FDLP encoded Hilbert envelopes) were substituted by white noise with uniform distribution. Such an algorithm provided subjectively much more natural signal than LPC10 standard (utilizing TDLP with order model equal to 10 estimated every 10ms) operating at twice higher bit-rates [6].

For NB speech applications operating at 8 kHz input signal, the sub-band residuals were split into equal length partially (5% - to avoid transient noise) overlapping segments. Each segment was heterodyned to DC range and Fourier transformed to yield spectral components of low-passed sub-band residuals. Commensurate number of spectral components in each sub-band was selected (using psychoacoustic model) and their parameters were vector quantized. In the decoder, the sub-band residuals were reconstructed and modulated with corresponding FDLP envelope. Individual DCT contributions from each critical sub-band were summed and inverse DCT was applied to reconstruct output signal [25].

IV. FROM NB TO WB-FDLP AUDIO CODEC

The first experiment towards wide-band (WB)-FDLP audio coding (48 kHz input signal) was motivated by the structure of the NB-FDLP speech codec operating at 8 kHz. The initial frequency sub-band decomposition based on weighting of DCT transformed signal was extended (by adding more critical sub-bands) to encode wide-band input [26].

In [27], a more efficient FDLP based version for WB audio coding was introduced. Initial critical bandwidth sub-band decomposition was replaced by quadrature mirror filter (QMF) bank. Then, FDLP was applied directly on QMF sub-band signals. Similar to the previous schemes, the sub-band FDLP residuals (the carrier signals for the FDLP-encoded Hilbert

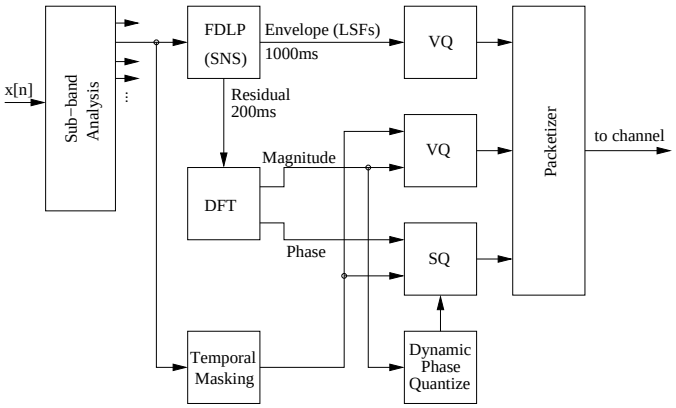


Fig. 2. Graphical scheme of the WB-FDLP encoder.

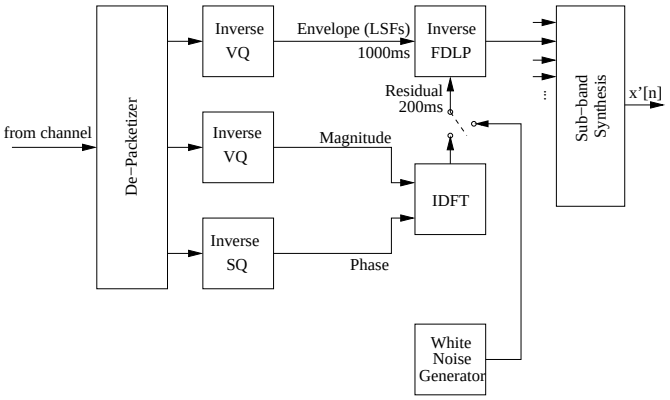


Fig. 3. Graphical scheme of the WB-FDLP decoder.

envelope) were further processed (more detailed description is given in the following Section IV-A). This WB-FDLP audio coding approach exploiting QMF bank decomposition serves as a simplified version (base-line system) of the current WB-FDLP audio codec.

The WB-FDLP approach is further improved by several additional blocks, described later, to operate at 66 kbps for audio signals sampled at 48 kHz. The encoder and decoder sides are described in the following sections.

A. Encoder

The block diagram of the WB-FDLP encoder is shown in Figure 2. On the encoder side, the full-band input signal is decomposed into QMF sub-bands. In each sub-band, FDLP technique is applied to approximate relatively long temporal sub-band envelopes (1000 ms). Resulting line spectral frequencies (LSFs) [28] approximating the sub-band temporal envelopes are quantized using split vector quantization (VQ) and selected codebook indices are transmitted to the decoder. Order of AR models is equal to 40. This number is a result obtained from optimization experiments, not reported here. The codebook used to quantize LSFs is trained across all QMF sub-bands using a set of audio-samples different from those used for objective and subjective quality evaluations.

LSFs are restored back at the encoder side and resulting AR model computed from quantized parameters is used to derive FDLP sub-band residuals (analysis-by-synthesis). Due to this operation, the quantization noise present in the sub-band temporal envelopes does not influence the reconstructed signal quality.

The sub-band FDLP residuals are derived by filtering the sub-band signal by the inverse FDLP filter. In order to take into account non-stationarity of the Hilbert carrier, FDLP residuals are split into 210 ms long sub-segments with 10 ms overlap. This ensures smooth transitions when the sub-segments of the residual signal are concatenated in the decoder. Each sub-segment is transformed into DFT domain. Magnitude spectral parameters are quantized using VQ. Phase spectral components of sub-band residuals are scalar quantized (SQ). In general, the number of levels in quantization differs for different sub-bands. Lower frequency sub-bands are quantized more accurately whereas the higher sub-bands exploit alternative techniques to reduce the overall bit-rates. More specifically:

- **Magnitudes:** Since a full-search VQ in this high dimensional space would be computationally demanding, split VQ approach is employed. Although the split VQ approach is suboptimal, it reduces computational complexity and memory requirements to manageable limits without severely affecting the VQ performance. We divide the input vector of spectral magnitudes into separate partitions of a lower dimension. Dimension of individual partitions varies with the frequency sub-band. In sub-bands $1-10$, the input vector of spectral magnitudes is split into 26 partitions (minimum codebook length is 3). Spectral magnitudes in higher sub-bands are quantized less accurately, i.e., the VQ codebook lengths increase. In overall, the VQ codebooks are trained (on a large audio database) for each partition using the LBG algorithm.
- **Phases:** The distribution of the phase spectral components was found to be approximately uniform (having a high entropy). Their correlation across time is not significant. Hence a uniform SQ is employed for encoding the phase spectral components. The SQ resolution varies from 3 to 5 bits, depending on the energy levels given by the corresponding spectral magnitudes, and is performed by a technique called dynamic phase quantization (DPQ). DPQ is described in details in Section V-B.

To reduce bit-rates, we apply an additional block, namely temporal masking (TM) which, together with DPQ, can efficiently control process of quantization. This block is described in Section V-E. Furthermore, a technique called spectral noise shaping (SNS, Section V-D) is applied for improving the quality of tonal signals by applying a TDLP filter prior to the FDLP processing. Detection of tonality followed by SNS is performed in each frequency sub-band independently.

B. Decoder

The block diagram of the WB-FDLP decoder is shown in Figure 3. On the decoder side, sub-band residuals are reconstructed by inverse quantization and are then modulated by temporal envelope given by FDLP model. FDLP model parameters are obtained from quantized LSFs.

More specifically, the transmitted VQ codebook indices are used to select appropriate codebook vectors for the magnitude spectral components. 210 ms segments of the sub-band residuals are restored in the time domain from its spectral magnitude and phase information. Overlap-add (OLA) technique is applied to obtain 1000 ms sub-band residuals, which are then modulated by the FDLP envelope to obtain the reconstructed sub-band signal.

An additional step of bit-rate reduction is performed on the decoder side (see Section V-C). FDLP residuals in frequency sub-bands above 12 kHz are not transmitted, but they are substituted by white noise at the decoder. Subsequently, these residuals are modulated by corresponding sub-band FDLP envelopes.

Finally, a block of QMF synthesis is applied on the reconstructed sub-band signals to produce the output full-band signal.

V. Individual blocks in WB-FDLP audio codec

This section describes, in detail, the major blocks employed in WB-FDLP audio codec mentioned in Section IV.

A. Non-uniform sub-band decomposition

The original sub-band decomposition in NB-FDLP speech codec was based on weighting of DCT sequence estimated from long-term full-band input signal by set of Gaussian windows [6]. In order to obtain non-uniform frequency sub-bands, Gaussian windows were distributed in a non-uniform way following the Bark warping function

$$z = 6sinh^{-1}(f/600), \tag{9}$$

where f and z are frequency axes in Hertz and in bark, respectively.

A higher efficiency is achieved by replacing the original sub-band decomposition by non-uniform QMF bank [29]. QMF provides the sub-band sequences which form a critically sampled and maximally decimated signal representation (i.e., the number of sub-band samples is equal to the number of input samples). In non-uniform QMF, the input audio (sampled at 48 kHz) is split into 1000 ms long frames. Each frame is decomposed into 32 non-uniform sub-bands. An initial decomposition with a 6 stage tree-structured uniform QMF analysis gives 64 uniformly spaced sub-bands. A non-uniform QMF decomposition into 32 frequency sub-bands is obtained by merging these 64 uniform QMF sub-bands [30]. This tying operation is motivated by critical band decomposition in the human auditory system. This means that more sub-bands at higher frequencies are merged together while maintaining perfect reconstruction. Magnitude frequency responses of first four QMF filters are given in Figure 4.

Unlike NB-FDLP coder, DCT is applied on the 1000 ms long sub-band signal to obtain AR model in a given QMF sub-band. STLP technique (introduced in Equation 8) is used to control the fit of AR model.

Such non-uniform QMF decomposition provides good compromise between fine spectral resolution for low frequency sub-bands and smaller number of FDLP parameters for higher bands. Furthermore, non-uniform QMF decomposition fits well into the perceptual audio coding scheme, where psychoacoustic models traditionally work in non-uniform (critical) sub-bands.

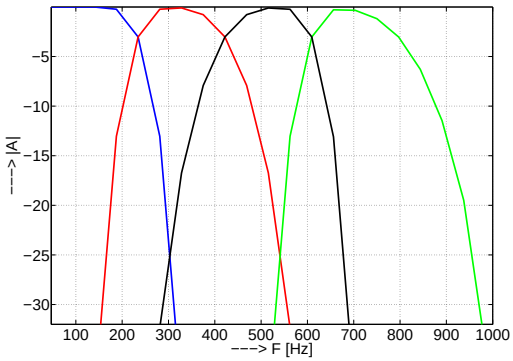


Fig. 4. Magnitude frequency response of first four QMF bank filters (filter length $N = 99$).

B. Dynamic phase quantization (DPQ)

To reduce bit-rate for representing phase spectral components of the sub-band FDLP residuals, we perform DPQ. DPQ can be seen as a special case of magnitude-phase polar quantization applied to audio coding [31].

In DPQ, graphically shown in Figure 5, phase spectral components corresponding to relatively low magnitude spectral components are transmitted with lower resolution, i.e., the codebook vector selected from the magnitude codebook is processed by "adaptive thresholding" in the encoder as well as in the decoder [25]. The threshold determines the resolution of quantization levels in uniform SQ. The threshold is dynamically adapted to meet a required number of phase spectral components for a given resolution. For frequency sub-band below 4 kHz, phase spectral components corresponding to the highest magnitudes are quantized with 5 bits, those corresponding to the lowest are quantized with 3 bits. For frequency sub-bands above 4 kHz, the highest resolution is 4 bits.

As DPQ follows an analysis-by-synthesis (AbS) scheme, no side information needs to be transmitted. This means that frequency positions of the phase components being dynamically quantized using different resolution do not have to be transmitted. Such information is available at the decoder side due to the perfect (lossless) reconstruction of magnitude components processed by AbS scheme.

C. White noise substitution

The detailed analysis of sub-band FDLP residuals shows that FDLP residuals from low frequency sub-bands resemble FM modulated signals. However, in high frequency sub-bands, the FDLP residuals have properties of white noise. According to these findings, we substitute FDLP residuals in frequency sub-bands above 12 kHz (last 3 bands) by white noise generated at the decoder side. These white noise residuals are then

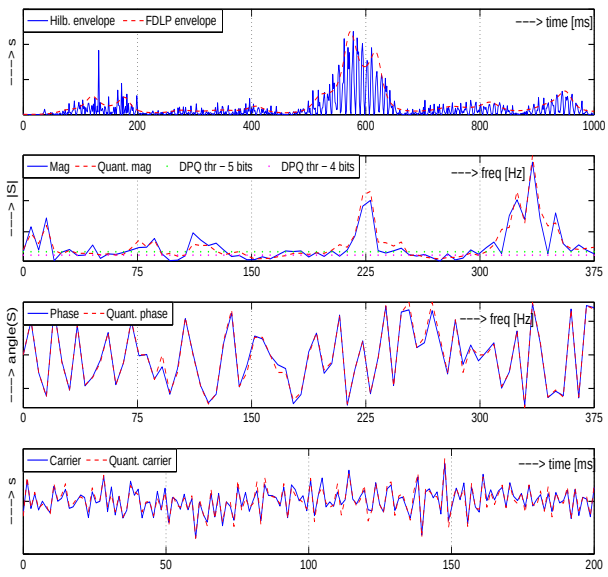


Fig. 5. Time-frequency characteristics obtained from a randomly selected audio example: (a) 1000 ms segment of the Hilbert envelope (estimated from the squared magnitude of an analytic signal) computed for the 3rd QMF sub-band, and its FDLP approximation. (b) Magnitude Fourier spectral components of the 200 ms sub-segment of the sub-band FDLP residual signal and its reconstructed version, adaptive thresholds for DPQ (5, 4 bits) are also shown. (c) Phase Fourier spectral components of the 200 ms sub-segment of the sub-band FDLP residual signal and its reconstructed version. (d) Original 200 ms sub-segment of the sub-band FDLP residual signal and its reconstructed version.

modulated by corresponding sub-band FDLP envelopes. White noise substitution of high sub-band residuals has a minimum impact on the quality of reconstructed audio (even for tonal signals) while providing a significant bit-rate reduction.

D. Spectral noise shaping (SNS)

The FDLP codec is most suitable to encode signals, such as glockenspiel, having impulsive temporal content, i.e., signals whose sub-band instantaneous energies can be characterized by an AR model. Therefore, FDLP is robust to "pre-echo" [7], [26] (i.e., quantization noise is spread before the onsets of the signal and may even exceed the original signal components in level during certain time intervals). However, for signals having impulsive spectral content, such as tonal signals, FDLP modeling approach is not appropriate. Here, most of the important signal information is present in the FDLP residual. For such signals, the quantization error in the FDLP codec spreads across all the frequencies around the tone. This results in significant degradation in the reconstructed signal quality.

This can be seen as the dual problem to encoding transients in the time domain, as done in many conventional codecs such as [3]. This is efficiently solved by temporal noise shaping (TNS) [7]. Specifically, coding artifacts arise mainly in handling transient signals (like the castanets) and pitched signals. Using spectral signal decomposition for quantization and encoding implies that a quantization error introduced in this domain will spread out in time after reconstruction by the synthesis filter bank. This phenomenon is called

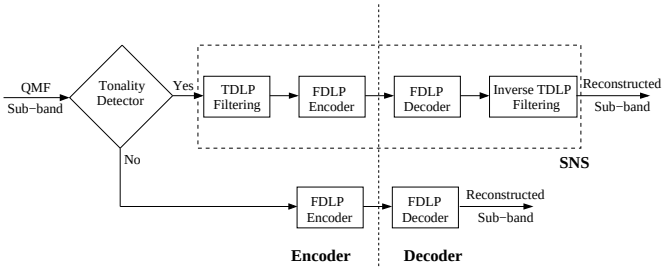


Fig. 6. WB-FDLP codec with SNS.

time/frequency uncertainty principle [32] and can cause "pre-echo" artifacts (i.e. a short noise-like event preceding a signal onset) which can be easily perceived. TNS represents one solution to overcome this problem by shaping the quantization noise in the time domain according to the input transient.

The proposed WB-FDLP audio codec exploits SNS technique to overcome problems in encoding tonal signals [33]. It is based on the fact that tonal signals are highly predictable in the time domain. If a sub-band signal is found to be tonal, it is analyzed using TDLP [20] and the residual of this operation is processed with the FDLP codec. At the decoder, the output of the FDLP codec is filtered by the inverse TDLP filter.

Since the inverse TDLP (AR) filter follows the spectral impulses for tonal signals, it shapes the quantization noise according to the input signal. General scheme of SNS module employed in WB-FDLP codec is given in Figure 6. SNS module consists of two blocks:

- Tonality detector (TD): TD identifies the QMF sub-band signals which have strong tonal components. Since FDLP performs well on non-tonal and partially tonal signals, TD ensures that only pure tonal signals are identified. For this purpose, global tonality detector (GTD) and local tonality detector (LTD) measures are computed and the tonality decision is taken based on both these measures. GTD measure is based on the spectral flatness measure (SFM, defined as the ratio of the geometric mean to the arithmetic mean of the spectral magnitudes) of the full-band signal. If the SFM is below the threshold, i.e., GTD has identified input frame as tonal, LTD is employed. LTD is defined based on the spectral auto-correlation of the sub-band signal (used for estimation of FDLP envelopes).
- SNS processing: If GTD and LTD have identified a sub-band signal to have a tonal character, such sub-band signal is filtered through the TDLP filter followed by FDLP model. Model orders of both models are equal to 20 as compared to a FDLP model order of 40 for the non-tonal signals. At the decoder side, inverse TDLP filtering on the FDLP decoded signal gives the sub-band signal back.

Improvements in reconstruction quality can be seen in Figure 7. For time-domain predicted signals, its TDLP filter has magnitude response characteristics similar to the power spectral density (PSD) of the input signal. As an example, Figure 8 shows the power spectrum of a tonal sub-band signal and the frequency response of the TDLP filter for this sub-

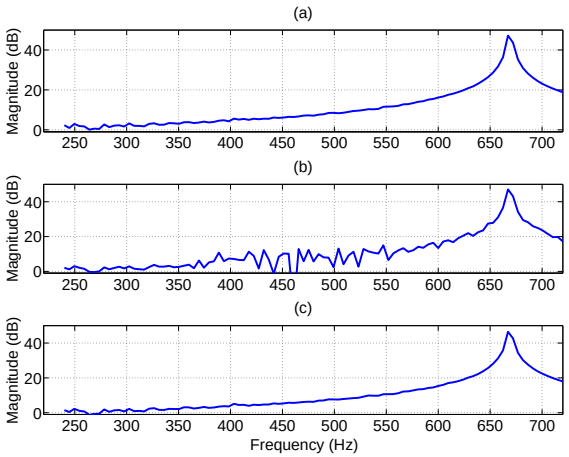


Fig. 7. Improvements in reconstruction signal quality with SNS: A portion of power spectrum of (a) a tonal input signal, (b) reconstructed signal using the WB-FDLP codec without SNS, and (c) reconstructed signal using the WB-FDLP codec with SNS.

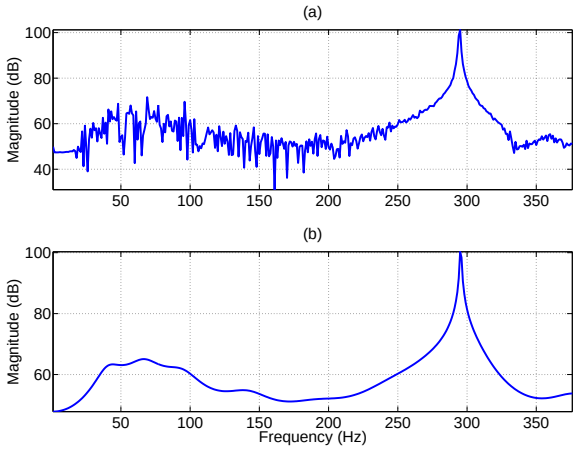


Fig. 8. TDLP filter used for SNS: (a) Power spectrum of tonal sub-band signal, and (b) magnitude response of the inverse TDLP filter in SNS.

band signal. Since the quantization noise passes through the inverse TDLP filter, it gets shaped in the frequency domain according to PSD of the input signal.

E. Temporal masking (TM)

A perceptual model, which performs temporal masking, is applied in WB-FDLP codec to reduce bit-rates. Temporal masking is a property of the human ear, where the sounds appearing within a temporal interval of about 200 ms after a signal component get masked. Such auditory masking property provides an efficient solution for quantization of a signal.

By processing relatively long temporal segments in frequency sub-bands, the FDLP audio codec allows for a straightforward exploitation of the temporal masking, while its implementation in more conventional short-term spectra based codecs has been so far quite limited, one notable exemption being the recently proposed wavelet-based codec [34].

The amount of forward masking is determined by the interaction of a number of factors including masker level, the temporal separation of the masker and the signal, frequency of the masker and the signal, and duration of the masker and the signal. We exploit linear forward masking model proposed in [35] to the sub-band FDLP residual signals. More particularly, a simple first order mathematical model, which provides a sufficient approximation for the amount of temporal masking is used

$$M[n] = a(b - \log_{10}\Delta t)(X[n] - c), \tag{10}$$

where M is the temporal mask in dB sound pressure level (SPL), X is the signal in dB SPL, n is the sample index, Δt is the time delay in ms, a, b and c are the constants. At any sample point, multiple mask estimates arising from the several previous samples are present and the maximum of it is chosen as the mask in dB SPL at that point. The optimal values of these parameters, as defined in [34], are as follows:

$$a = k_2 f^2 + k_1 f + k_0, \tag{11}$$

where f is the center frequency of the sub-band in kHz, k_0, k_1 and k_2 are constants. The constant b is obtained from the duration of the temporal masking and may be chosen as $\log_{10} 200$. The parameter c is the absolute threshold of hearing (ATH) in quiet, defined as

$$c = 3.64 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 0.001 f^4. \tag{12}$$

To estimate the masking threshold at each sample index, we compute a short-term dB SPL so that the signal is divided into 10 ms overlapping frames with frame shifts of 1 sample.

The assumptions made in the applied linear model, such as sinusoidal nature of the masker and signal, minimum duration of the masker (300 ms), minimum duration of the signal (20 ms) may differ from real audio signal encoding conditions. Therefore, the actual masking thresholds are much below the thresholds obtained from the linear masking model. To obtain the actual thresholds, informal listening experiments were conducted to determine the correction factors [36].

These masking thresholds are then utilized in quantizing the sub-band FDLP residual signals. The number of bits required for representing the sub-band FDLP residuals is reduced in accordance with TM thresholds compared to the WB-FDLP codec without TM. Since the sub-band signal is the product of its FDLP envelope and residual (carrier), the masking thresholds for the residual signal are obtained by subtracting the dB SPL of the envelope from that of the sub-band signal. First, we estimate the quantization noise present in the WB-FDLP codec without TM. If the mean of the quantization noise (in 210 ms sub-band signal) is above the masking threshold, no bit-rate reduction is applied. If the temporal mask mean is above the noise mean, then the amount of bits needed to encode that sub-band FDLP residual signal is reduced in such a way that the noise level becomes similar to the masking threshold.

An example of application TM is shown in Figure 9. We plot a region of the 200 ms sub-band signal, the quantization noise before and after applying TM. As can be seen, in some regions, the instantaneous quantization noise levels present in the FDLP codec after applying TM can be slightly higher

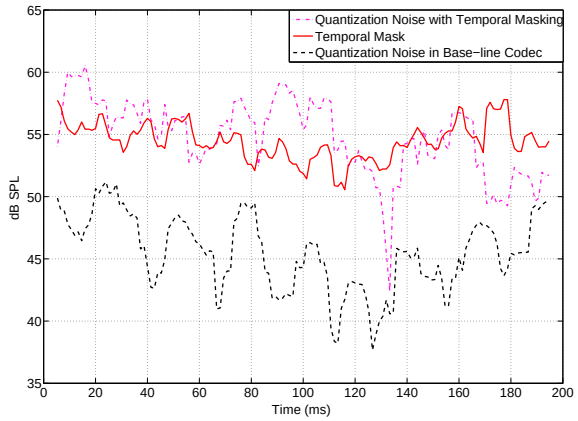


Fig. 9. Application of temporal masking (TM) to reduce the bits for 200 ms region of a high energy sub-band signal. The figure shows the temporal masking threshold for a high-energy region of sub-band signal, quantization noise for the WB-FDLP codec without TM and for the codec with TM.

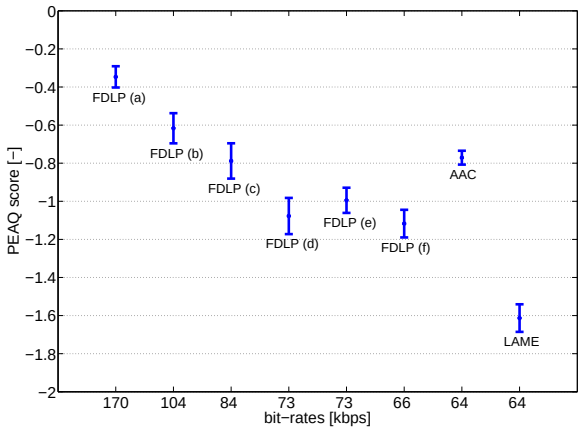


Fig. 10. PEAQ (ODG) scores: mean values and 95% confidence intervals estimated over 27 audio recordings to evaluate individual blocks of WB-FDLP codec (see Section VI). We add results for MPEG-4 HE-AAC and LAME-MP3 codecs. PEAQ (ODG) score meanings are given in Table II.

than corresponding TM thresholds. However, the mean of the quantization noise is smaller than the mean of TM threshold over the whole 200 ms long time segment. Since the information regarding the number of quantization bits needs to be transmitted to the receiver, the bit-rate reduction is done in a discretized manner. Due to that, the quantization noise needs to be only roughly estimated (over the whole segment) and the mean value is compared to the mean TM threshold. Specifically, in the WB-FDLP codec, the bit-rate reduction is done in 8 different levels (in which the first level corresponds to no bit-rate reduction).

VI. Bit-rate versus quality of the individual blocks

To evaluate individual blocks employed in WB-FDLP codec, we perform quality assessment and provide achieved results with obtained bit-rate reductions. For the quality assessment, perceptual evaluation of audio quality (PEAQ) distortion measure [37] is used. PEAQ measure, based on the ITU-R BS.1387 standard, estimates the perceptual degradation of the test signal with respect to the reference signal. The output combines a number of model output variables (MOV's) into a single measure, the objective difference grade (ODG) score, which is an impairment scale with meanings shown in Table II.

PEAQ evaluations are performed on 27 challenging audio recordings sampled at 48 kHz. These audio samples form part of the MPEG framework for exploration of speech and audio coding [38]. They are comprised of speech, music and speech over music recordings, and specifically mentioned in Table I.

Mean values and 95% confidence intervals of ODG scores obtained by PEAQ measure for 27 audio samples are shown in Figure 10 for the various WB-FDLP codec versions:

(a) Base-line system (170 kbps): This version of the codec employs uniform QMF decomposition and DFT magnitudes of sub-band residuals are quantized using split VQ. Quantization of the spectral magnitudes using the split VQ allocates about 30 kbps for all the frequency sub-bands. DFT phases are uniformly quantized using 5 bits. Such codec operates at 170 kbps [27].

(b) Non-uniform QMF decomposition (104 kbps): Employment of non-uniform QMF bank decomposition (Section V-A) significantly reduces the bit-rates from 170 kbps to 104 kbps (about 40%), while the overall objective quality is degraded by about 0.25.

(c) Dynamic phase quantization (DPQ) (84 kbps): Employment of DPQ (Section V-B) provides bit-rate reduction about 20 kbps (from 104 to 84 kbps), while the overall objective quality is degraded by about 0.1.

(d) Noise substitution (73 kbps): Subsequent white noise substitution of high frequency sub-band FDLP residuals (Section V-C) reduces the bit-rate to 73 kbps (by about 11 kbps), while the overall objective quality is degraded by about 0.3.

(e) Spectral noise shaping (SNS) (73 kbps): SNS block employed to improve encoding of highly tonal signals (Section V-D) increases bit-rates by 32 bps (to transmit the binary decision about employment of SNS in each sub-band). SNS does not affect the encoding of non-tonal signals. Overall quality was slightly improved by about 0.1 [33].

For purpose of detailed evaluation, 5 additional test signals with strong tonality structure, downloaded from [39], were used in the experiments. Due to the application of SNS, the objective quality of each of these recordings is improved, as shown in Figure 11. The average objective quality score (average PEAQ score) for these samples is improved by about 0.4.

(f) Temporal masking (TM) (66 kbps): Final block simulates temporal masking to modify quantization levels of spectral components of sub-band residuals according to perceptual significance (Section V-E). The bit-rate reduction is about 7 kbps for an average PEAQ degradation by about 0.1 [36].

VII. Comparison with state-of-the-art audio codecs

The final version of the WB-FDLP codec operating at 66 kbps, which employs all the blocks described and evaluated in

	test recordings
Speech	chinese female, es02[1,2], es03, louis raquin[1], te19
Music	brahms, dongwoo[1,2], es01[2], phi1, phi2[2] - phi_3[1] - phi7, salvation, sc03[2], te09[1], te15[2], trilogy[1]
Speech and music	Arirang, Green, Wedding, te1_mg54
Speech over music	noodleking[1,2], te16_fe49, twinkle[1]

TABLE I

List of 27 audio/speech recordings selected for objective quality assessment. [1] denotes 8 recordings used in MUSHRA subjective listening test. [2] denotes 5 recordings used in BS.1116 subjective listening test.

ODG Scores	Quality
0	imperceptible
−1	perceptible but not annoying
−2	slightly annoying
−3	annoying
−4	very annoying

TABLE II

PEAQ (ODG) scores and their meanings.

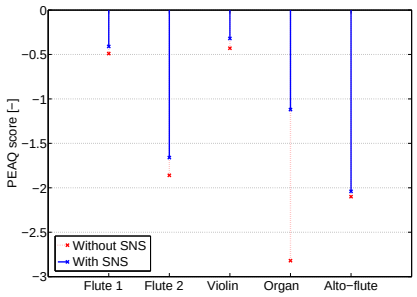


Fig. 11. PEAQ (ODG) scores for 5 selected tonal files encoded by WB-FDLP codec (at 66 kbps) with and without SNS. PEAQ (ODG) score meanings are given in Table II.

Sections V and VI, respectively, is compared with the state-of-the-art MPEG audio codecs. In our evaluations, the following two codecs are considered:

1) LAME - MP3 (3.97 32bits) (MPEG 1, layer 3) at 64 kbps [40]. Lame codec based on MPEG-1 architecture [2] is currently considered the best MP3 encoder at mid-high bit-rates and at variable bit-rates.
2) MPEG-4 HE-AAC (V8.0.3), v1 at 64 kbps [3]. The HE-AAC coder is the combination of spectral band replication (SBR) [41] and advanced audio coding (AAC) [42] and was standardized as high-efficiency AAC (HE-AAC) in Extension 1 of MPEG-4 Audio [43].

Objective quality evaluation results for the 27 speech/audio files from [38] encoded by LAME-MP3 and MPEG-4 HE-AAC codecs are also present in Figure 10.

Subjective evaluation of the proposed WB-FDLP codec with respect to two state-of-the-art codecs (LAME-MP3 and MPEG4 HE-AAC) is carried out by MUSHRA (multiple stimuli with hidden reference and anchor) methodology. It is defined by ITU-R recommendation BS.1534 [44]. We perform the MUSHRA tests on 8 audio samples from the database [38] with 22 listeners. The recordings (originals as well as encoded versions) selected for MUSHRA evaluations, specifically mentioned in Table I, can be downloaded from [45]. The results of the MUSHRA tests are shown in Figure 12. In Figure 13, we also show MUSHRA test results for 4 expert listeners for

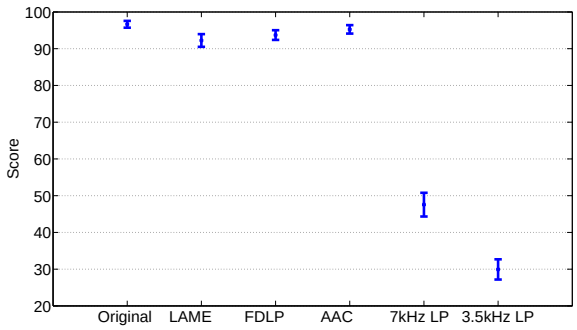


Fig. 12. MUSHRA results for 8 audio files with 22 listeners encoded using three codecs: WB-FDLP (66 kbps), MPEG-4 HE-AAC (64 kbps) and LAME-MP3 (64 kbps). We add results for hidden reference (original) and two anchors (7 kHz low-pass filtered and 3.5 kHz low-pass filtered). We show mean values and 95% confidence intervals.

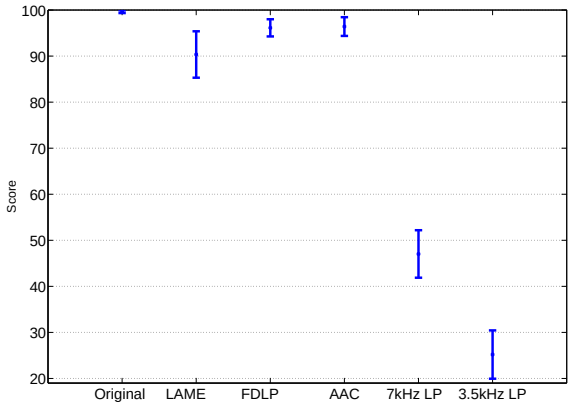


Fig. 13. MUSHRA results for 8 audio files with 4 expert listeners encoded using three codecs: WB-FDLP (66 kbps), MPEG-4 HE-AAC (64 kbps) and LAME-MP3 (64 kbps). We add results for hidden reference (original) and two anchors (7 kHz low-pass filtered and 3.5 kHz low-pass filtered). We show mean values and 95% confidence intervals.

the same data. These listeners are included in the previous list of 22 subjects.

Furthermore, in order to better understand the performances of the proposed WB-FDLP codec, we perform the BS.1116 methodology of subjective evaluation [46]. BS.1116 is used to detect small impairments of the encoded audio compared to the original. As this subjective evaluation is time consuming, only two coded versions (proposed WB-FDLP and MPEG-4 HE-AAC) are compared. The subjective results with 7 listeners using 5 speech/audio samples from the same database [38], mentioned in Table I, are shown in Figure 14.

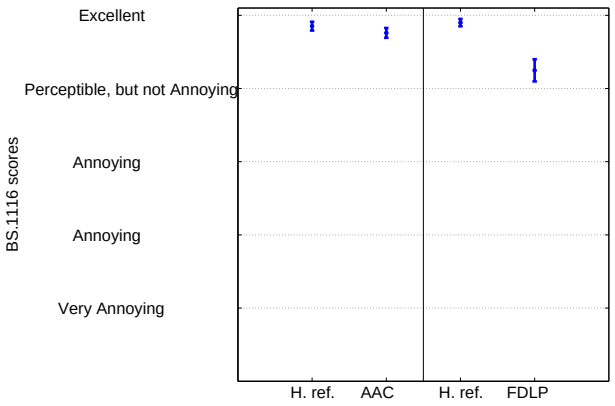


Fig. 14. BS.1116 results for 5 audio files with 7 listeners encoded using two codecs: WB-FDLP (66 kbps), MPEG-4 HE-AAC (64 kbps). We add results for hidden reference (H. ref.) for each codec separately. We show mean values and 95% confidence intervals.

VIII. Discussions and conclusions

A novel wide-band audio compression system for medium bit-rates is presented. The audio codec is based on processing relatively long temporal segments of the input audio signal. Frequency domain linear prediction (FDLP) is applied to exploit predictability of temporal evolution of spectral energies in non-uniform sub-bands of the signal. This yields sub-band residuals, which are quantized using temporal masking. The use of FDLP ensures that fine temporal details of the signal envelopes are captured with high temporal resolution. Several additional techniques are used to reduce the final bit-rate. The proposed compression system is relatively simple and suitable for coding both speech and music.

Performances of the some of the individual processing steps are evaluated using objective perceptual evaluation of audio quality, standardized by ITU-R (BS.1387). Final performances of the codec at 66 kbps are evaluated using subjective quality evaluation (MUSHRA and BS.1116 standardized by ITU-R). The subjective evaluation results suggest that the proposed WB-FDLP codec provides better audio quality than LAME - MP3 codec at 64 kbps and produces slightly worse results compared to MPEG-4 HE-AAC standard at 64 kbps.

We stress that the codec processes each frequency sub-band independently without taking into account sub-band correlations, which could further reduce the bit-rate. This strategy has been pursued intentionally to ensure robustness to packet losses. The drop-out of bit-packets in the proposed codec corresponds to loss of sub-band signals at the decoder. In [47], it has been shown that the degraded sub-band signals can be efficiently recovered from the adjacent sub-bands in time-frequency plane which are unaffected by the channel.

From computational complexity point of view, the proposed codec does not perform highly demanding operations. Linear prediction coefficients of the FDLP model are estimated using fast LBG algorithm. Most of the computational cost is due to the search of appropriate codewords to vector quantize magnitude spectral components of the sub-band residuals. However, codebook search limitations, which also applied in traditional speech codecs such as CELP, have been already overcome by various techniques (e.g. two-stage algebraic-stochastic quantization scheme).

The fundamental technique - FDLP - used in the presented audio codec is a frequency-domain dual of the well-known time-domain linear prediction (TDLP). Similar to this duality, also the other techniques exploited in the proposed codec can be associated to standard signal processing techniques:

- QMF: it performs frequency-domain alias cancellation. This is a dual property to a technique called time-domain alias cancellation (TDAC). TDAC ensures perfect invertibility of the modified discrete cosine transform (MDCT) used in AAC codecs.
- SNS: Unlike temporal noise shaping (TNS) employed in AAC codecs to outperform problems with transient signals, SNS improves quality of highly tonal signals compressed by the WB-FDLP codec.
- TM: it is a psychoacoustic phenomenon implemented to significantly reduce bit-rates while maintaining the quality of the reconstructed audio. TM is often referred to as non-simultaneous masking (part of auditory masking), where sudden stimulus sound makes inaudible other sounds which are present immediately preceding or following the stimulus. Since the effectiveness of TM lasts approximately 100ms (in case of the offset attenuation), TM is a powerful and easily implementable technique in the FDLP codec. Frequency masking (FM) (or simultaneous masking) is a dual phenomenon to TM, where a sound is made inaudible by a “masker”, a noise of the same duration as the original sound. FM is exploited in most of psychoacoustic models used by traditional audio codecs.

Modern audio codecs combine some of the previously mentioned dual techniques (e.g., QMF and MDT implemented in adaptive transform acoustic coding (ATRAC) developed by Sony [48]) to improve perceptual qualities/bit-rates. Due to this, we believe that there is still a potential to improve the efficiency of the FDLP codec that has not been pursued yet. For instance, the proposed version of the codec does not utilize standard entropy coding. Further, neither SNRs in the individual sub-bands are evaluated nor signal dependent non-uniform quantization in different frequency sub-bands (e.g. module of frequency masking discussed above) and at different time instants (e.g. bit-reservoir) are employed. Inclusion of these techniques should further reduce the required bit-rates and provide bit-rate scalability, which form part of our future work.

Acknowledgment

This work was partially supported by grants from ICSI Berkeley, USA; the Swiss National Center of Competence in Research (NCCR) on “Interactive Multi-modal Information Management (IM)2”; managed by the IDIAP Research Institute on behalf of the Swiss Federal Authorities. The authors would like to thank Vijay Ullal and Marios Athineos for their active involvement in the development of the codec. They would also like to thank the reviewers for providing numerous helpful comments on the manuscript.

REFERENCES

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," *in Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 10, pp. 937-940, Tampa, USA, April 1985,

[2] K. Brandenburg, et al., "The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio," *in 92nd Convention of Audio Engineering Society (AES)*, preprint 3336, New York, USA, 1992.

[3] J. Herre, J, M. Dietz, "MPEG-4 high-efficiency AAC coding", *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 137 - 142, May 2008.

[4] M. S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," *in Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, pp. 3277-3280, Salt Lake City, USA, April 2001.

[5] H. Dudley, "The carrier nature of speech," *Bell System Technical Journal*, vol. 19, no. 4, pp. 495-515, October 1940.

[6] P. Motlicek, H. Hermansky, H. Garudadri, and N. Srinivasamurthy, "Speech Coding Based on Spectral Dynamics," *in Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, DE, pp. 471-478, September 2006.

[7] J. Herre, and J. H. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," *in 101st Convention of Audio Engineering Society (AES)*, preprint 4384, November 1996.

[8] M. Athineos, and D. P. W. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," *in Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pp. 648651, Hong Kong, April 2003.

[9] R. Kumerasan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *Journal of Acoustical Society of America*, vol. 105, no 3, pp. 1912-1924, March 1999.

[10] S. Ganapathy, P. Motlicek, H. Hermansky and H. Garudadri, "Autoregressive Modeling of Hilbert Envelopes for Wide-band Audio Coding," in 124th Convention of Audio Engineering Society (AES), Amsterdam, Netherlands, May 2008.

[11] M. G. Christensen, and S. H. Jensen, "Computationally Efficient Amplitude Modulated Sinusoidal Audio Coding Using Frequency-Domain Linear Prediction," *in Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pp. 189-192, Toulouse, France, May 2006.

[12] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *in Rep. 6th International Congr. Acoustic*, Y. Kohasi Ed., pp. C17-C20, Paper C-5-5, August 1968.

[13] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell System Technical Journal*, vol. 49, no. 6, pp. 1973-1986, October 1970.

[14] S. Ganapathy, T. Samuel, H. Hermansky, "Modulation Frequency Features For Phoneme Recognition In Noisy Speech", J. Acoust. Soc. Am., Express letters, January 2009.

[15] T. Houtgast, H.J.M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function, I. General room acoustics," *Acustica 46*, pp. 60-72, 1980.

[16] IEC 60268-16: "Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index", *<http://www.iec.ch/>*

[17] B.E.D. Kingsbury, N. Morgan, S. Greenberg, "Robust speech recognition using the modulation spectrogram", *Speech Communication*, Vol. 25 , Issue 1-3, pp. 117-132, August 1998.

[18] M. Athineos, H. Hermansky, and D. P. W. Ellis, "LP-TRAP: Linear predictive temporal patterns," *in Proc. of ICSLP*, pp. 1154-1157, Jeju, S. Korea, October 2004.

[19] T. H. Falk, S. Stadler, W. B. Kleijn and W. Chan, "Noise Suppression Based on Extending a Speech-Dominated Modulation Band," *Interspeech 2007*, pp. 970-973, Antwerp, Belgium, August 2007.

[20] J. Makhoul, "Linear Prediction: A Tutorial Review," *in Proc. of IEEE*, Vol. 63, No. 4, April 1975.

[21] M. Athineos, D. P. W. Ellis, "Autoregressive modeling of temporal envelopes," *in IEEE Trans. Speech and Audio Processing*, vol. 55, pp. 5237-5245, November 2007.

[22] A. V. Oppenheim, and R. W. Schafer, "Discrete-Time Signal Processing," 2nd Ed., Prentice-Hall, NJ, USA, 1998.

[23] R. V. Churchill, and . W. Brown, "Introduction to Complex Variables Applications," 5th Ed., McGraw-Hill Book Company, NY, USA, 1982.

[24] H. Hermansky, H. Fujisaki, and Y. Sato, "Analysis and Synthesis of Speech based on Spectral Transform Linear Predictive Method," *in Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 8, pp. 777-780, Boston, USA, April 1983.

[25] P. Motlicek, H. Hermansky, S. Ganapathy, and H. Garudadri, "Non-Uniform Speech/Audio Coding Exploiting Predictability of Temporal Evolution of Spectral Envelopes," *in Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, DE, pp. 350-357, September 2007.

[26] P. Motlicek, V. Ullal, and H. Hermansky, "Wide-Band Perceptual Audio Coding Based on Frequency-Domain Linear Prediction," *in Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pp. 265-268, Honolulu, USA, April 2007.

[27] P. Motlicek, S. Ganapathy, H. Hermansky, and H. Garudadri, "Frequency Domain Linear Prediction for QMF Sub-bands and Applications to Audio coding," *in Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, DE, pp. 248-258, June 2007.

[28] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals", J. Acoust. Soc. Am., vol. 57, S35, 1975.

[29] A. Charbonnier, and J-B Rault, "Design of nearly perfect non-uniform QMF filter banks," *in Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1786-1789, New York, NY, USA, April 1988.

[30] P. Motlicek, S. Ganapathy, H. Hermansky, H. Garudadri and M. Athineos, "Perceptually motivated Sub-band Decomposition for FDLP Audio Coding," *in Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, DE, pp. 435-442, September 2008.

[31] R. Vafin, and W. B. Kleijn,"Entropy-constrained polar quantization and its applications to audio coding," *in IEEE Trans. Speech and Audio Processing*, vol. 13, pp. 220-232, March 2005.

[32] M. A. Pinsky, "Introduction to Fourier Analysis and Wavelets," *Pacific Grove, CA : Brooks/Cole*, 2002.

[33] S. Ganapathy, P. Motlicek, H. Hermansky, H. Garudadri, "Spectral Noise Shaping: Improvements in Speech/Audio Codec Based on Linear Prediction in Spectral Domain," *in Proc. of Interspeech*, Brisbane, Australia, September 2008.

[34] F. Sinaga, T. S. Gunawan and E. Ambikairajah, "Wavelet Packet Based Audio Coding Using Temporal Masking," *IEEE conference on Information, Communications and Signal Processing 2003*, pp. 1380-1383, Singapore, December 2003.

[35] W. Jesteadt, S. P. Bacon, and J. R. Lehman, "Forward Masking as a function of frequency, masker level, and signal delay," J. Audio Eng. Soc., Vol. 71(4), pp. 950-962, April 1982.

[36] S. Ganapathy, P. Motlicek, H. Hermansky, and H. Garudadri, "Temporal Masking for Bit-rate Reduction in Audio Codec Based on Frequency Domain Linear Prediction," *in Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4781 - 4784, Las Vegas, April 2008.

[37] ITU-R Recommendation BS.1387, "Method for objective psychoacoustic model based on PEAQ to perceptual audio measurements of perceived audio quality," December 1998.

[38] ISO/IEC JTC1/SC29/WG11: "Framework for Exploration of Speech and Audio Coding," MPEG2007/N9254, Lausanne, Switzerland, July 2007.

[39] Musical instrumental samples: *<http://theremin.music.uiowa.edu/MIS.html>*.

[40] LAME-MP3 codec: *<http://lame.sourceforge.net>*.

[41] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral Band Replication, a novel approach in audio coding," *in 112th Convention of Audio Engineering Society (AES)*, preprint 5553, Munich, DE, May 2002.

[42] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," J. Audio Eng. Soc., vol. 45, no. 10, pp. 789-814, October 1997.

[43] ISO/IEC, "Coding of audio-visual objects Part 3: Audio, AMENDMENT 1: Bandwidth Extension," ISO/IEC Int. Std. 14496-3:2001/Amd.1:2003, 2003.

[44] ITU-R Recommendation BS.1534: "Method for the subjective assessment of intermediate audio quality," June 2001.

[45] Homepage with encoded samples: *<http://www.idiap.ch/∼pmotlic>*.

[46] ITU-R Recommendation BS.1116: "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," October 1997.

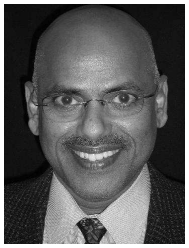
[47] S. Ganapathy, P. Motlicek, and H. Hermansky, "Error Resilient Speech Coding Using Sub-band Hilbert Envelopes," *in Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, DE, pp. 355-362, September 2009.

[48] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, and R.M. Heddle, "ATRAC: Adaptive Transform Acoustic Coding for MiniDisc," in 93rd Convention of Audio Engineering Society (AES), preprint 3456, San Francisco, USA, October 1992.

Petr Motlicek received the M.Sc. degree in electrical engineering and the Ph.D. degree in computer science from Brno University of Technology (BUT), Czech Republic, in 1999 and 2003, respectively. In 2000 he worked on very low bit-rate speech coding at École Supérieure d´Ingénieurs en Électrotechnique et Électronique (ESIEE), Paris, France. From 2001 to 2002, he worked as a Research Assistant at Oregon Graduate Institute (OGI), Portland, USA, in the area of distributed speech recognition. Since 2005, he has been a Research Scientist at Idiap Research Institute, Martigny, Switzerland. His research interests include speech processing, feature extraction for robust automatic speech recognition and speech and audio coding. From 2000, Dr. Motlicek is a member of IEEE and ISCA. From 2004, he holds a position of Assistant Professor in the speech processing group at BUT.

Sriram Ganapathy Sriram Ganapathy received his B.Tech degree in Electronics and Communication from University of Kerala, India in 2004 and the M.E. degree in Signal Processing from Indian Institute of Science, Bangalore in 2006. From 2006 to 2008, he worked as a Research Assistant in Idiap Research Institute, Switzerland. He is currently pursuing his Ph.D. at Center for Language and Speech Processing, Dept. of ECE, Johns Hopkins University, USA. His research interests include signal processing, audio coding and robust speech recognition.

Hynek Hermansky is a Full Professor of the Electrical and Computer Engineering at the Johns Hopkins University in Baltimore, Maryland. He is also a Professor at the Brno University of Technology, Czech Republic, an Adjunct Professor at the Oregon Health and Sciences University, Portland, Oregon, and an External Fellow at the International Computer Science Institute at Berkeley, California. He is a Fellow of IEEE for Invention and development of perceptually-based speech processing methods, was the Technical Chair at the 1998 ICASSP in Seattle and an Associate Editor for IEEE Transaction on Speech and Audio. Further he is Member of the Editorial Board of Speech Communication, holds 6 US patents and authored or co-authored over 200 papers in reviewed journals and conference proceedings. He has been working in speech processing for over 30 years, previously as a Director of Research at the IDIAP Research Institute, Martigny, and an Adjunct Professor at the Swiss Federal Institute of Technology in Lausanne, Switzerland, a Professor and Director of the Center for Information Processing at OHSU Portland, Oregon, a Senior Member of Research Staff at U S WEST Advanced Technologies in Boulder, Colorado, a Research Engineer at Panasonic Technologies in Santa Barbara, California, and a Research Fellow at the University of Tokyo. He holds Dr. Eng. Degree from the University of Tokyo, and Dipl. Ing. Degree from Brno University of Technology, Czech Republic. His main research interests are in acoustic processing for speech recognition.



Harinath (Hari) Garudadri has been with Qualcomm, San Diego, CA, U.S.A. Since 1997. His current interests include biomedical signal processing and low power sensors for telemetry of vital signs for wireless health and fitness applications. He worked on Audio Coding, Voice Recognition, Speech Coding, Video Coding, Packet Switched Video Telephony and Multimedia Standards at Qualcomm. Hari has over 30 peer-reviewed publications and over 30 issued/pending patents in these areas. Prior to Qualcomm, Hari worked at Voice Processing Corporation, Cambridge, MA; INRS Telecommunications, Montreal, Quebec; Microtel Pacific Research, Burnaby, B.C.; and Indian Institute of Technology, Kanpur. Hari has a Ph.D in EE from University of British Columbia (1988) and M.Tech from Indian Institute of Technology, Mumbai (1980).